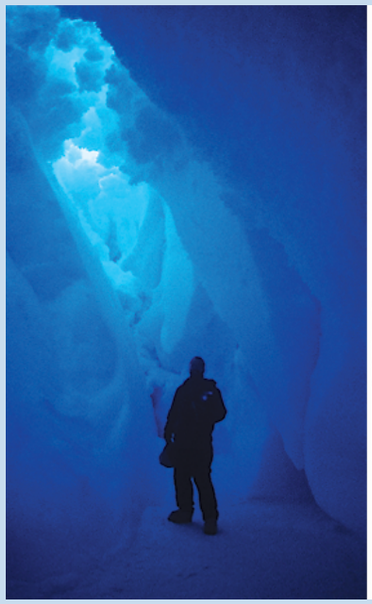
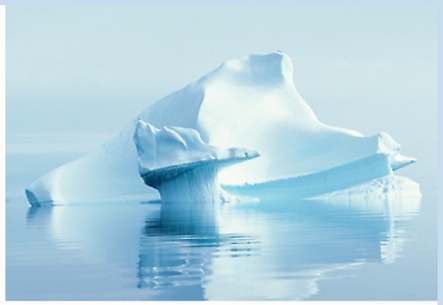
National Snow and Ice Data Center

Supporting Cryospheric Research Since 1976





Identifying Data in the Earth Sciences

R. Duerr (and a cast of many)

Outline

- Background
- What do we mean by identifiers for data?
- Practical and impractical use cases
- Identifier schemes and assessment criteria
- Results and recommendations
- Towards Best Practices

Home — Technology Infusion Working Group

home | Federation of Earth Science Information Partners

http://www.esipfed.org/

Most Visited | Getting Started | Elevations | WorldMark | Igloo | Shove this | TinyURL! | home page | INAK | ESIP wiki | DC wiki | DC – JIRA

home | Federation of Earth Scie... | NASA – Crews and Expeditions

Federation of Earth Science Information Partners

MAKING DATA MATTER

Home | About | Membership | Meetings | Community | Resources | Opportunities | News

ESIP

- About
- News
- Membership
- Partners
- Meetings
- Community Collaborations
- Resources
- Opportunities
- Wiki

Portals

- GCMD ESIP Data Portal
- GCMD ESIP Services Portal

Member Login

The Federation of Earth Science Information Partners

The ESIP Federation is a diverse network of scientists, data stewards and technology developers that:

- Improves access to Earth science data and information.
- Connects public, academic and private providers to each other and users of Earth science data and information.
- Promotes consensus-based solutions and best practices affecting the Earth science data and information community
- Provides a neutral forum for Earth science data experts to collaborate, learn and solve communitywide problems affecting access, dissemination and use of Earth science data and information.

Learn More...

Meetings

The 2011 ESIP Federation Winter Meeting will be held January 4-6 in Washington, DC.

- Details
- Registration

Newsletter

The ESIP Federation's newsletter is published bimonthly in February, April, June, August, October and December. Any submission by the beginning of the month is welcome.

Click here for the latest issue.

Click here to submit news.

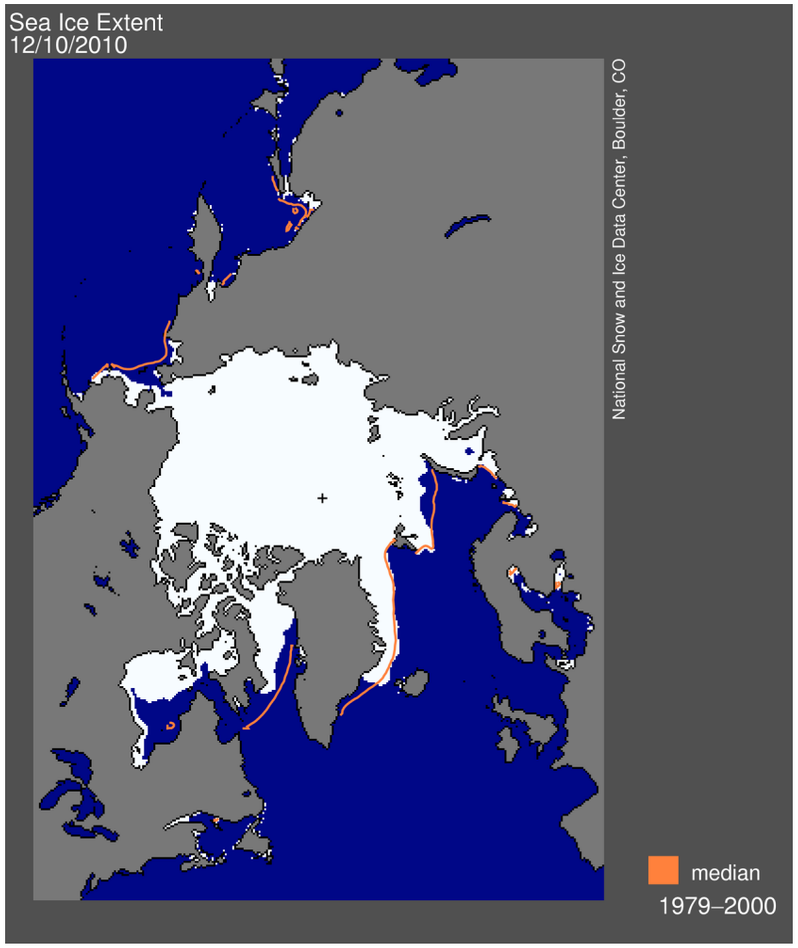
Calendar

http://www.esipfed.org/resources

NSIDC National Snow and Ice Data Center

What do we mean by identifiers for data?

- A unique name for these data
- The location of that named data today or any day in the future



Data Levels

- Only two levels of data identifier were addressed:
  - A collection or data set as a whole
  - Individual items (files or granules) within a data set

Use Case #1: Unique Identifier

- To uniquely & unambiguously identify a digital object no matter which copy a user has
- Ideal attributes
  - Location independent (I.e., copies everywhere have this same ID)
  - Generate at time of object creation
  - Placeable inside the object or it's metadata
- Practical attributes
  - Globally unique
  - No name authority
  - Relatively difficult to change
- Write once and don't maintain model

Use Case #2: Unique Locator

- To locate a copy of the digital object no matter where it is currently held
- Ideal attributes
  - Location invariant (I.e., no matter where the object moves, this ID remains the same and can always be used to find it)
  - Globally unique
- Practical attributes
  - External name authority necessary
  - Generate only on decision to make data permanently available
- Maintain forever model

Use Case #3: Citable Identifier

- To identify data cited in a particular publication
- Ideal attributes
  - Basically those of a Unique Locator with a couple of caveats
    - Acceptance by publishers and authors
    - Facilitate identification at the data set or data set subset level
    - Granule level citation not practical in most cases at the current time

Use Case #4: Scientifically Unique Identifier

- To be able to tell that two digital objects contain the same data even if the formats are different.

- Ideal Attributes

  - Same as Unique Identifier plus

  - Possible to verify that the contents are unchanged after a format transformation or certain kinds of content rearrangement

Identifier schemes assessed

- Archival Resource Key (ARK)
- Digital Object Identifiers (DOI)
- Extensible Resource Identifier (XRI)
- HANDLE
- Life Science ID (LSID)
- Object Identifiers (OID)
- Persistent Uniform Resource Locators (PURL)
- URI/URN/URL
- UUID

Assessment Criteria

- Technical value (Standard? Security? Scalability? Interoperability? Internet compatibility? 3rd party maintenance? Naming authority and stability? Expected longevity?)
- User value (Usable in citations? Any additional trust value? Opaque or transparent?)
- Archive value (Costs, Ease of migration, Extensible to non-web based objects, physical objects?)
- Existing usage within data centers

Assessment Results: Use Cases

	Unique Identifier		Unique Locator		Citable Locator		Scientifically Unique Identifier	
ID Scheme	**Dataset**	**Item**	**Dataset**	**Item**	**Dataset**	**Item**	**Dataset**	**Item**
ARK								
DOI								
XRI								
Handle								
LSID								
OID								
PURL								
URL/URN/ URI								
UUID								

Best Practices

- Recognize that different identifier schemes are meant to solve different problems
- Recognize that a minimum of two identifiers will be needed for any data set or data file
- Plan for scheme obsolescence and replacement

Recommendations

- Assign UUIDs for each data file or granule in your data sets
- Assign a DOI for each data set so that they may be cited

Next Steps

- UUID for granules/files and DOI for data sets will be submitted to SPG for potential endorsement as NASA standards
- Identifiers paper to be submitted to Journal of Earth Science Informatics
- Preservation and Stewardship cluster have agreed to work on recommendations for citations for both data users and data producers/archives over the next year
- ESIP Preservation and Stewardship cluster identifier testbed activities continue in the hopes that practical experience may bring further clarity